**Documentação Técnica – Projeto Data Lake “Faixa de Risco”**

**1. IDENTIFICAÇÃO DO GRUPO**

* **Nome do Projeto:** Faixa de Risco: Onde os dados cruzam com o perigo
* **Setor de Atuação:** Setor Público (Trânsito)
* **Integrantes :**

| **Nome completo** | **E-mail** | **LinkedIn** |
| --- | --- | --- |
| **Guilherme Predolin Dino** | [**Guilherme\_predolin@hotmail.com**](mailto:Guilherme_predolin@hotmail.com) | [**Link**](https://www.linkedin.com/in/guilherme-predolin-dino-636489122/) |
| **Carolina Althman** | **Carolina.althman@gmail.com** | [**Link**](https://www.linkedin.com/in/calthman/) |
| **Ítalo Rufca Catelli Fernandez** | **Italo.rufca@gmail.com** | [**Link**](https://www.linkedin.com/in/italo-rufca/) |

**2. DESCRIÇÃO DO PROBLEMA DE NEGÓCIO**

**2.1 Qual é o problema que vocês pretendem resolver?**

Entender as bases de dados de infrações/acidentes em rodovias federais e mapear os principais focos de acidentes e a probabilidade de os mesmos acontecerem.

**2.2 Qual é o impacto do problema para o negócio?**

Diminuir a quantidade de acidentes/infrações e, em consequência disso, salvar vidas.

**3. OBJETIVOS DO PROJETO**

**3.1 Objetivo Geral:**

Construir um pipeline escalável e confiável para ingestão, tratamento e análise de dados de infrações e acidentes rodoviários no Brasil, visando subsidiar decisões de políticas públicas e prevenção de acidentes.

**3.2 Objetivos Específicos (mínimo 3):**

* Análise dos Dados
* Análise Exploratória
* Aplicar Modelo de Regressão: Extra Trees Regressor

**4. Arquitetura da Solução**

**Escopo:**

* Ingestão de arquivos CSV de acidentes disponibilizados pelo governo
* Armazenamento em camadas Raw, Trusted e Refined
* Normalização, padronização de dados e enriquecimento com base de dados de clima e feriados
* Disponibilização para dashboards analíticos (Tableau)

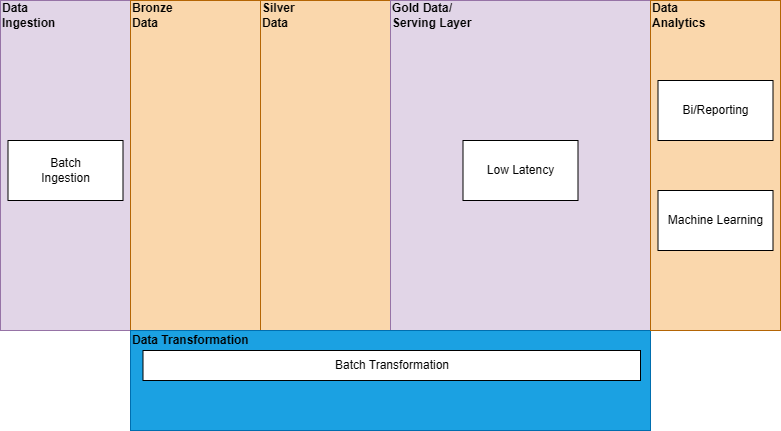
**Camadas do Data Lake:**

* **Raw**: Armazenamento bruto dos dados originais (.csv)
* **Trusted**: Dados validados, limpos e normalizados
* **Refined**: Dados prontos para análise, agregados e otimizados

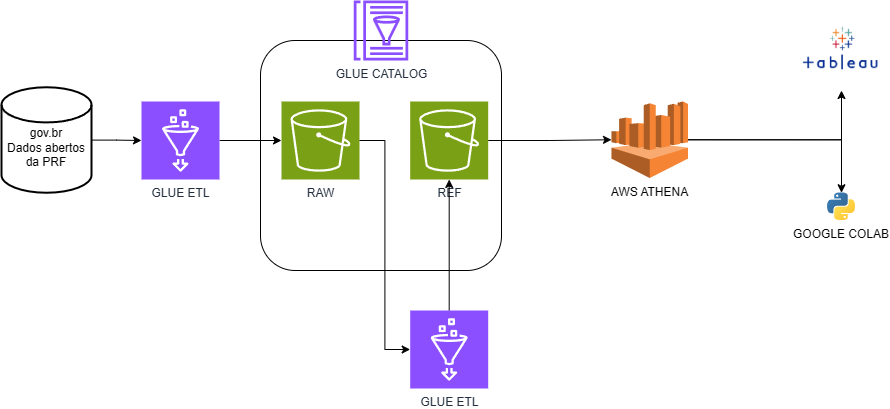
**Serviços AWS Utilizados:**

| **Componente** | **Serviço AWS** | **Função** |
| --- | --- | --- |
| Armazenamento | S3 | Camadas Raw, Trusted, Refined |
| Processamento | AWS Glue (Python shell) | Scripts ETL e transformação de arquivos |
| Orquestração | Step Functions | Controle de fluxo ETL e agendamento |
| Catálogo de Dados | AWS Glue Data Catalog | Metadados e tabelas externas para Athena |
| Consulta e BI | Athena / Tableau | Exploração e visualização dos dados |

**Arquitetura Conceitual**



**Arquitetura Lógica**



**5. Pipeline de Dados**

# 5.1 Ingestão

* Fonte: [Portal de Dados Abertos - PRF](https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf)
* Frequência: Mensal
* Formato: CSV com encoding ISO-8859-1
* Armazenamento Inicial:
  + s3://handson-fund-dados-analytics/raw/raw\_acidentes/
  + s3://handson-fund-dados-analytics/raw/raw-calendario/

s3://handson-fund-dados-analytics/raw/raw-infracoes/

**Script de Ingestão**

import os

import boto3

# Configurações do S3

s3 = boto3.client('s3')

raw\_bucket = ' handson-fund-dados-analytics'

raw\_prefix = ' raw/raw\_acidentes/' # prefixo dentro do bucket

# Caminho local dos arquivos que deseja enviar

local\_folder = 'C:\Users\Guilh\Documents'

def upload\_folder\_to\_s3(local\_path, bucket, prefix):

for root, dirs, files in os.walk(local\_path):

for file in files:

local\_file\_path = os.path.join(root, file)

# Montar o caminho relativo para manter a estrutura no S3

relative\_path = os.path.relpath(local\_file\_path, local\_path)

s3\_key = os.path.join(prefix, relative\_path).replace("\\", "/")

try:

s3.upload\_file(local\_file\_path, bucket, s3\_key)

print(f"✅ Enviado: {local\_file\_path} -> s3://{bucket}/{s3\_key}")

except Exception as e:

print(f"❌ Erro ao enviar {local\_file\_path}: {e}")

# Execução

if \_\_name\_\_ == "\_\_main\_\_":

upload\_folder\_to\_s3(local\_folder, raw\_bucket, raw\_prefix)

# 5.2 Processamento e Transformação

| **Etapa** | **Descrição** |
| --- | --- |
| Validação de Schema | Checagem de colunas esperadas, tipos de dados e valores nulos |
| Conversão de encoding | ISO-8859-1 → UTF-8 |
| Conversão de formatos | CSV → Parquet (melhor performance para Athena) |
| Enriquecimento | Junção com base de feriados nacionais e previsão do tempo via API externa |
| Padronização de datas | dd/mm/yyyy → yyyy-mm-dd |
| Normalização de UF | Letras maiúsculas e validação contra lista IBGE |

# 5.3 Camadas e Tabelas

# Camadas

# Camada RAW

### **raw\_acidentes**

• Finalidade: Armazenar os dados brutos de acidentes de trânsito oriundos do gov.br (PRF).

• Formato de arquivo: CSV com delimitador \;

• Localização no S3: s3://handson-fund-dados-analytics/raw/raw-acidentes/

• Características:

- Estrutura de dados bruta, sem normalização.

- Contém colunas como data\_inversa, uf, tipo\_acidente, latitude, entre outras.

- Utiliza skip.header.line.count = 1 para ignorar cabeçalho.

### **raw\_infracoes**

• Finalidade: Armazenar dados brutos de infrações de trânsito.

• Formato de arquivo: CSV com delimitador \;

• Localização no S3: s3://handson-fund-dados-analytics/raw/raw-infracoes

• Características:

- Informações como numero\_auto, data\_infracao, codigo\_infracao, descricao\_tipo\_veiculo.

### **raw\_calendario**

• Finalidade: Armazenar dados de datas, feriados e atributos temporais.

• Formato de arquivo: CSV com delimitador ,

• Localização no S3: s3://handson-fund-dados-analytics/raw/raw-calendario

# Camada Refined

### **ref\_acidentes**

• Finalidade: Estruturação dos dados de acidentes após tratamento.

• Transformações aplicadas:

- Conversão de data\_inversa para tipo date

- Correções de encoding

- Padronização dos campos

* • Localização: s3://handson-fund-dados-analytics/refined/ref-acidentes

### **ref\_infracoes**

• Finalidade: Versão tratada da tabela de infrações.

• Transformações aplicadas:

- Conversão de campos de data (data\_infracao, inicio\_vigencia\_infracao) para date

• Localização: s3://handson-fund-dados-analytics/raw/raw-infracoes

### **ref\_calendario**

• Finalidade: Versão tratada da tabela de calendário com cast de campos e tratamento de datas.

• Localização: s3://handson-fund-dados-analytics/refined/ref-calendario

# Camada Specialized

### **spec\_acidentes**

• Finalidade: Enriquecimento dos dados de acidentes com o contexto do calendário.

• Processamento:

- CTE cte\_acidentes extrai dados da ref\_acidentes

- CTE cte\_calendario inclui campo derivado tag\_feriado

- LEFT JOIN realizado pela coluna data\_inversa = data

• Criação: CREATE TABLE AS SELECT

### **spec\_infracoes**

• Finalidade: Enriquecimento dos dados de infrações com dados do calendário

• Processamento:

- CTE cte\_infracoes extrai dados da ref\_infracoes

- CTE cte\_calendario aplica classificação de feriado

- LEFT JOIN realizado pela coluna data\_infracao = data

Considerações Finais

• Todas as tabelas RAW são EXTERNAL, garantindo que os dados não sejam apagados com drop da tabela.

• As camadas Refined aplicam tratamentos de qualidade e conformidade com tipos adequados.

• As tabelas Specialized permitem análises enriquecidas para BI ou exploração em notebooks.

• O pipeline respeita boas práticas de separação de responsabilidades entre ingestão, tratamento e consumo.

## Criação de Tabelas

# Camada RAW

### **raw\_acidentes**

DDL SQL para criação da tabela:

CREATE EXTERNAL TABLE raw\_handson.raw\_acidentes(  
 id bigint,   
 pesid string,   
 data\_inversa string,   
 dia\_semana string,   
 horario string,   
 uf string,   
 br bigint,   
 km string,   
 municipio string,   
 causa\_principal string,   
 causa\_acidente string,   
 ordem\_tipo\_acidente bigint,   
 tipo\_acidente string,   
 classificacao\_acidente string,   
 fase\_dia string,   
 sentido\_via string,   
 condicao\_metereologica string,   
 tipo\_pista string,   
 tracado\_via string,   
 uso\_solo string,   
 id\_veiculo string,   
 tipo\_veiculo string,   
 marca string,   
 ano\_fabricacao\_veiculo string,   
 tipo\_envolvido string,   
 estado\_fisico string,   
 idade string,   
 sexo string,   
 ilesos string,   
 feridos\_leves string,   
 feridos\_graves string,   
 mortos string,   
 latitude string,   
 longitude string,   
 regional string,   
 delegacia string,   
 uop string)  
ROW FORMAT DELIMITED   
 FIELDS TERMINATED BY '\;'   
STORED AS INPUTFORMAT   
 'org.apache.hadoop.mapred.TextInputFormat'   
OUTPUTFORMAT   
 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'  
LOCATION 's3://handson-fund-dados-analytics/raw/raw-acidentes/'  
TBLPROPERTIES ('skip.header.line.count'='1');

### **raw\_infracoes**

DDL SQL para criação da tabela:

CREATE EXTERNAL TABLE raw\_handson.raw\_infracoes(  
 numero\_auto string,   
 data\_infracao string,   
 indicador\_bordagem string,   
 assinatura\_auto string,   
 sentido\_trafego string,   
 uf\_infracao string,   
 br\_infracao bigint,   
 km\_infracao bigint,   
 municipio string,   
 indicador\_veiculo\_estrangeiro string,   
 uf\_placa string,   
 descricao\_especie\_veiculo string,   
 descricao\_marca\_veiculo string,   
 descricao\_tipo\_veiculo string,   
 descricao\_modelo\_veiculo string,   
 codigo\_infracao bigint,   
 descricao\_abreviada\_infracao string,   
 enquadramento\_infracao string,   
 inicio\_vigencia\_infracao string,   
 fim\_vigencia\_infracao string,   
 medicao\_infracao string,   
 hora\_infracao bigint,   
 medicao\_considerada bigint,   
 excesso\_verificado string,   
 qtd\_infracoes bigint)  
ROW FORMAT DELIMITED   
 FIELDS TERMINATED BY '\;'   
STORED AS INPUTFORMAT   
 'org.apache.hadoop.mapred.TextInputFormat'   
OUTPUTFORMAT   
 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'  
LOCATION 's3://handson-fund-dados-analytics/raw/raw-infracoes'  
TBLPROPERTIES ('skip.header.line.count'='1');

### **raw\_calendario**

DDL SQL para criação da tabela:

CREATE EXTERNAL TABLE raw\_handson.raw\_calendario(  
 data date,   
 dia bigint,   
 mes bigint,   
 ano bigint,   
 dia\_semana string,   
 feriado string)  
ROW FORMAT DELIMITED   
 FIELDS TERMINATED BY ','   
STORED AS INPUTFORMAT   
 'org.apache.hadoop.mapred.TextInputFormat'   
OUTPUTFORMAT   
 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'  
LOCATION 's3://handson-fund-dados-analytics/raw/raw-calendario'  
TBLPROPERTIES ('skip.header.line.count'='1');

### **Camada Refined**

### **ref\_acidentes**

DDL SQL para criação da tabela:

CREATE EXTERNAL TABLE refined\_handson.ref\_acidentes(  
 id bigint,   
 pesid string,   
 data\_inversa date,   
 dia\_semana string,   
 horario string,   
 uf string,   
 br bigint,   
 km string,   
 municipio string,   
 causa\_principal string,   
 causa\_acidente string,   
 ordem\_tipo\_acidente bigint,   
 tipo\_acidente string,   
 classificacao\_acidente string,   
 fase\_dia string,   
 sentido\_via string,   
 condicao\_metereologica string,   
 tipo\_pista string,   
 tracado\_via string,   
 uso\_solo string,   
 id\_veiculo string,   
 tipo\_veiculo string,   
 marca string,   
 ano\_fabricacao\_veiculo string,   
 tipo\_envolvido string,   
 estado\_fisico string,   
 idade string,   
 sexo string,   
 ilesos string,   
 feridos\_leves string,   
 feridos\_graves string,   
 mortos string,   
 latitude string,   
 longitude string,   
 regional string,   
 delegacia string,   
 uop string)  
ROW FORMAT DELIMITED   
 FIELDS TERMINATED BY '\;'   
STORED AS INPUTFORMAT   
 'org.apache.hadoop.mapred.TextInputFormat'   
OUTPUTFORMAT   
 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'  
LOCATION 's3://handson-fund-dados-analytics/refined/ref-acidentes'  
TBLPROPERTIES ('skip.header.line.count'='1');

### **ref\_infracoes**

DDL SQL para criação da tabela:

CREATE EXTERNAL TABLE refined\_handson.ref\_infracoes(  
 numero\_auto string,   
 data\_infracao date,   
 indicador\_bordagem string,   
 assinatura\_auto string,   
 sentido\_trafego string,   
 uf\_infracao string,   
 br\_infracao bigint,   
 km\_infracao bigint,   
 municipio string,   
 indicador\_veiculo\_estrangeiro string,   
 uf\_placa string,   
 descricao\_especie\_veiculo string,   
 descricao\_marca\_veiculo string,   
 descricao\_tipo\_veiculo string,   
 descricao\_modelo\_veiculo string,   
 codigo\_infracao bigint,   
 descricao\_abreviada\_infracao string,   
 enquadramento\_infracao string,   
 inicio\_vigencia\_infracao date,   
 fim\_vigencia\_infracao string,   
 medicao\_infracao string,   
 hora\_infracao bigint,   
 medicao\_considerada bigint,   
 excesso\_verificado string,   
 qtd\_infracoes bigint)  
ROW FORMAT DELIMITED   
 FIELDS TERMINATED BY '\;'   
STORED AS INPUTFORMAT   
 'org.apache.hadoop.mapred.TextInputFormat'   
OUTPUTFORMAT   
 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'  
LOCATION 's3://handson-fund-dados-analytics/raw/raw-infracoes'  
TBLPROPERTIES ('skip.header.line.count'='1');

### **ref\_calendario**

DDL SQL para criação da tabela:

CREATE EXTERNAL TABLE refined\_handson.ref\_calendario(  
 data date,   
 dia bigint,   
 mes bigint,   
 ano bigint,   
 dia\_semana string,   
 feriado string)  
ROW FORMAT DELIMITED   
 FIELDS TERMINATED BY ','   
STORED AS INPUTFORMAT   
 'org.apache.hadoop.mapred.TextInputFormat'   
OUTPUTFORMAT   
 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'  
LOCATION 's3://handson-fund-dados-analytics/refined/ref-calendario'  
TBLPROPERTIES ('skip.header.line.count'='1');

### **spec\_acidentes**

DDL SQL para criação da tabela:

create table specialized\_handson.spec\_acidentes as (

with cte\_calendario as (

SELECT

data

, dia

, mes

, ano

, dia\_semana

, feriado

, case

when feriado = '' then 'Não'

else 'Sim'

end as tag\_feriado

FROM "refined\_handson"."ref\_calendario"

),

cte\_acidentes as (

select

id

, pesid

, data\_inversa

, dia\_semana as dia\_semana\_acidente

, horario

, uf

, br

, km

, municipio

, causa\_principal

, causa\_acidente

, ordem\_tipo\_acidente

, tipo\_acidente

, classificacao\_acidente

, fase\_dia

, sentido\_via

, condicao\_metereologica

, tipo\_pista

, tracado\_via

, uso\_solo

, id\_veiculo

, tipo\_veiculo

, marca

, ano\_fabricacao\_veiculo

, tipo\_envolvido

, estado\_fisico

, idade

, sexo

, ilesos

, feridos\_leves

, feridos\_graves

, mortos

, latitude

, longitude

, regional

, delegacia

, uop

from refined\_handson.ref\_acidentes

)

select

\* from

cte\_acidentes a

LEFT JOIN cte\_calendario c

on a.data\_inversa = c.data

)

### **spec\_infracoes**

DDL SQL para criação da tabela:

create table specialized\_handson.spec\_infracoes as (

with cte\_calendario as (

SELECT

data

, dia

, mes

, ano

, dia\_semana

, feriado

, case

when feriado = '' then 'Não'

else 'Sim'

end as tag\_feriado

FROM "refined\_handson"."ref\_calendario"

),

cte\_infracoes as (

select

numero\_auto

, data\_infracao

, indicador\_bordagem

, assinatura\_auto

, sentido\_trafego

, uf\_infracao

, br\_infracao

, km\_infracao

, municipio

, indicador\_veiculo\_estrangeiro

, uf\_placa

, descricao\_especie\_veiculo

, descricao\_marca\_veiculo

, descricao\_tipo\_veiculo

, descricao\_modelo\_veiculo

, codigo\_infracao

, descricao\_abreviada\_infracao

, enquadramento\_infracao

, inicio\_vigencia\_infracao

, fim\_vigencia\_infracao

, medicao\_infracao

, hora\_infracao

, medicao\_considerada

, excesso\_verificado

, qtd\_infracoes

from refined\_handson.ref\_infracoes

)

select

\* from

cte\_infracoes i

LEFT JOIN cte\_calendario c

on i.data\_infracao = c.data

)

**6. Esquema de Dados**

| **Campo** | **Tipo** | **Descrição** |
| --- | --- | --- |
| id | bigint | Identificador único do acidente |
| data\_acidente | date | Data normalizada do acidente |
| uf | string | Unidade Federativa |
| municipio | string | Nome do município |
| tipo\_acidente | string | Classificação (ex: Colisão, Atropelamento) |
| clima | string | Condição climática no momento do acidente |
| feriado | string | Flag se era feriado nacional |
| br | string | Número correspondente à BR da infração/acidente |
| km | string | Número correspondente ao KM da infração/acidente |

**7. Segurança e Governança**

* Criptografia dos buckets (SSE-S3)
* Controle de acesso via IAM (principais roles: GlueJobRole)
* Versionamento de objetos habilitado no bucket
* Catalogação automática com Glue Crawlers (Trusted/Refined)

**8. Dashboard Final**

**Ferramenta:** Tableau  
**Insights:**

* Acidentes por BR e município
* Acidentes por tipo e horário
* Quantitativo de mortes e feridos
* Correlação entre clima e acidentes
* Evolução temporal por estado